Hybrid Legal Precedent Retrieval via LLM-Based Document Expansion and Dense-Sparse Fusion

Otávio Gabriel de Oliveira Santana[1], **Anderson Bessa da Costa**[1], **Edson Takashi Matsubara**[1]

[1]Faculty of Computing (FACOM)
Federal University of Mato Grosso do Sul (UFMS)
Campo Grande – MS – Brazil
{otavio.g, anderson.bessa, edson.matsubara}@ufms.br

Abstract. *Precedent retrieval in the Brazilian legal domain faces severe vocabulary mismatch between narrative queries and technical court decisions. This work investigates the Hypothetical Document Embeddings (HyDE) framework, a zero-shot semantic expansion approach, on legal search. We built a benchmark comprising structured appellate decisions from the State Court of Mato Grosso do Sul (TJMS), modeling narrative syllabi sections as queries and legal theses as targets. We evaluate HyDE, its hybridization with BM25 via Reciprocal Rank Fusion (RRF), and query granularity using the* `google/gemma-3-12b-it` *model. Experiments demonstrate that HyDE outperforms dense retrieval, quadrupling Recall@1 (from 0.086 to 0.345) and boosting MRR from 0.131 to 0.427 ($p < 0.001$). Furthermore, hybrid fusion (HyDE + BM25) reaches an MRR of 0.571, achieving statistical equivalence to pure lexical search (MRR = 0.574). Query granularity analysis indicates that medium-length queries (combining facts and legal controversy) represent the optimal operational point, maximizing precision without redundant overhead. Ultimately, the hybrid architecture emerges as a robust real-world solution, proving the viability of Large Language Models (LLMs) in modernizing specialized legal research.*

Keywords: *Legal Precedent Retrieval; Large Language Models; Generative Document Expansion; Dense-Sparse Fusion; HyDE.*

1. Introduction

Efficient legal precedent retrieval is a core task, as prior decisions substantiate arguments and guide statutory interpretation. The Brazilian context is marked by a volume of pending lawsuits exceeding 83 million [Conselho Nacional de Justiça 2024], making legal search systems critical. However, legal language features high terminological density and complex syntax [Chalkidis et al. 2020]. This hinders query-document matching, especially when queries are formulated as natural narratives rather than technical court decisions.

Traditionally, information retrieval systems rely on sparse lexical approaches, such as TF-IDF and BM25 [Robertson and Zaragoza 2009]. While effective, these methods struggle with semantic similarity in domains exhibiting high terminological variability. Recently, dense retrieval methods utilizing neural embeddings have demonstrated consistent gains by modeling semantic relationships [Karpukhin et al. 2020, Reimers and Gurevych 2019], outperforming purely lexical baselines.

Despite these advances, dense models rely on annotated data for fine-tuning. In the legal domain, this annotation requires domain experts, making labeled dataset creation an expensive and time-consuming process [Chalkidis et al. 2022]. In this context, zero-shot approaches emerge as promising alternatives. Notably, the HyDE (Hypothetical Document Embeddings) framework [Gao et al. 2023] leverages a Large Language Model (LLM) to generate a hypothetical document from the original query. In practice, the model infers expected technical terminology, bridging the user's narrative and the target document's vocabulary. The embedding of this generated text is then used as the query representation in the dense space.

Although sparse methods such as BM25 excel when exact lexical overlap exists, they struggle when users formulate queries as natural factual narratives without technical legal jargon. Conversely, dense models bridge this semantic gap but typically require expensive domain-specific fine-tuning. Therefore, our core hypothesis is that combining generative semantic expansion (HyDE) with sparse exact matching (BM25) provides a complementary effect: HyDE overcomes the vocabulary mismatch in exploratory searches, while BM25 anchors literal terms, yielding a superior and more resilient architecture than using either method in isolation.

To guide this research, we formulate three Research Questions (RQs):

- **RQ1 (Overall Effectiveness):** To what extent does generative semantic expansion (HyDE) improve the retrieval of Portuguese legal syllabi compared to traditional sparse and dense baselines?
- **RQ2 (Hybridization):** Does combining HyDE with hybrid retrieval strategies (Dense + Sparse) yield superior performance gains over using semantic expansion in isolation?
- **RQ3 (Query Granularity):** How do query length and specificity (e.g., isolated factual narratives *versus* the fully structured body) affect the effectiveness of HyDE?

To answer these questions, we built a curated corpus comprising appellate decisions from the State Court of Mato Grosso do Sul (TJMS)[1]. These documents reflect the modern standard of structured syllabi, distinctly divided into a Syllabus Header (thematic indexers and core legal thesis) and a Structured Body (factual narrative and legal controversy). In our setup, narrative sections of the Body serve as queries, while the corresponding Header acts as the target document. Evaluation is performed using classic metrics: Recall@k, MRR, and nDCG@k.

The main contributions of this work are summarized as follows:

- The creation and release of a controlled benchmark grounded in real-world structured data for evaluating legal information retrieval in Brazil;
- An empirical analysis of a zero-shot generative expansion approach (HyDE) operating in a high-terminological-density Portuguese scenario;
- A systematic evaluation of hybrid search architectures (dense-sparse) utilizing LLM-expanded representations;

[1]Available at: https://esaj.tjms.jus.br/cjsg/consultaCompleta.do. Accessed on: June 13, 2026.

- An investigation into the impact of query length and granularity on generation quality.

2. Related Work

This section reviews the information retrieval literature, emphasizing dense approaches, hybridization strategies, semantic expansion via language models, and applications in the legal domain.

Information Retrieval: Lexical, Dense, and Hybrid Approaches. Traditional sparse methods such as BM25 excel at exact matching [Robertson and Zaragoza 2009] but struggle with semantic similarity in domains with high terminological variability, suffering from vocabulary mismatch. Dense retrieval mitigates this by projecting texts into continuous vector spaces, capturing semantics without lexical overlap. While models such as Sentence-BERT [Reimers and Gurevych 2019] and Dense Passage Retrieval (DPR) [Karpukhin et al. 2020] popularized neural embeddings, recent models such as the E5 family [Wang et al. 2022] support zero-shot multilingual scenarios crucial for Portuguese. However, dense methods remain highly sensitive to query quality. To address this, hybrid approaches combine sparse and dense methods to merge exact matching, vital for legal jargon, with deep semantic similarity. Unsupervised fusion strategies such as Reciprocal Rank Fusion (RRF) [Cormack et al. 2009] integrate these signals without fine-tuning, consistently outperforming standalone methods [Lin et al. 2022].

Semantic Query Expansion via Language Models. Classic expansion strategies reduce semantic mismatch [Carpineto and Romano 2012], but Large Language Models (LLMs) now enable generative enrichment. While Retrieval-Augmented Generation (RAG) [Lewis et al. 2020] grounds generation on retrieved texts, recent literature explores the inverse: using generation to enhance retrieval. A prominent example is HyDE [Gao et al. 2023], a zero-shot approach where an LLM generates a synthetic target document that is embedded to capture dense relevance. Internationally, integrating these generative techniques into legal NLP applications and retrieval pipelines has shown promising results in tackling domain-specific challenges [Ariai et al. 2025]. We hypothesize that this generative expansion is uniquely suited for Portuguese legal retrieval to resolve severe vocabulary mismatch between narrative facts and abstract court theses, especially when fused with sparse lexical signals to preserve statutory exactness.

Natural Language Processing in the Legal Domain. NLP in the legal domain spans classification and extraction tasks [Chalkidis et al. 2020], with international benchmarks such as LexGLUE [Chalkidis et al. 2022] and COLIEE [Rabelo et al. 2022] driving neural and hybrid retrieval advancements. In Brazil, adaptation to Portuguese has progressed through foundational datasets such as the VICTOR corpus [Luz de Araujo et al. 2020] for document classification. However, legal retrieval faces unique vocabulary mismatch challenges. Early Brazilian efforts relied on non-scalable symbolic AI and Case-Based Reasoning [Bueno et al. 1999]. Recently, the Ulysses project evaluated dense and sparse (BM25) retrieval on audit documents [Fernandes et al. 2026], proving the efficacy of exact lexical matching in Portuguese. Despite these advances, international reviews highlight limits for

neural retrieval in complex statutory domains [Ariai et al. 2025, Lai et al. 2024], a gap exacerbated in Brazil by the scarcity of systematic studies on domestic court particularities [Oliveira et al. 2024].

3. Methodology

This section presents the formal problem modeling, describes the structure and criteria of the legal corpus, and details the strategies for vector representation, query expansion, and hybrid retrieval, concluding with the evaluation protocol.

3.1. Problem Formulation and Corpus

Let $D = \{d_1, \ldots, d_{501}\}$ be a legal corpus of public appellate decisions collected via random sampling from the State Court of Mato Grosso do Sul (TJMS) public repository, issued between 2003 and 2026. Included documents meet strict criteria: a complete structured syllabus, a minimum of 500 words, full-text availability, no judicial secrecy, and uniqueness across legal proceedings (ensuring no duplicate appellate records). Quantitatively, the Header and Body average 64 and 657 words, respectively. The dataset covers 14 procedural classes, predominantly Civil Procedure, mainly comprising Civil (229) and Criminal (144) Appeals.

Following court standardization, each decision d_i natively comprises two distinct textual components:

- **Syllabus Header** (c_i): consolidates thematic indexers and the established legal thesis. In our experiment, this acts as the target document (Key).
- **Structured Body** (s_i): details the case and legal controversy narratively. This acts as the Query.

Table 1 illustrates a translated corpus example.

Formally, given queries $Q = \{q_1, \ldots, q_{501}\}$ derived from the Structured Bodies (s_i), the system must locate the exact corresponding Header c_i within the target collection $C = \{c_1, \ldots, c_{501}\}$. Although q_i and c_i originate from the same decision, they represent distinct communicative layers. The descriptive natural language of the Body and the highly abstract terminology of the Header inherently generate a severe vocabulary mismatch. Mapping the narrative body to its corresponding header accurately simulates real-world challenges where practitioners describe a factual case to find relevant jurisprudence. To ensure reproducibility, the dataset and mappings are publicly available[2].

3.2. Vector Representation and Search Strategies

For retrieval, texts are mapped into a continuous vector space. Let $f(\cdot)$ be an open-source multilingual neural embedding model that projects a text into $\mathbb{R}^k$. The vector index is built by mapping all target headers in the collection:

$$e_j = f(c_j), \quad j = 1, \ldots, N \tag{1}$$

[2]Available at: https://github.com/otvio-olive/hyde-legal-retrieval-dataset. Accessed on: June 21, 2026

Table 1. Example of a structured document in the corpus (Translated for illustrative purposes)

Field	Content
Class/Subject	Civil Appeal / ICMS
Origin	State Court of Mato Grosso do Sul (TJMS) — 2025
Header (c_i)	TAX LAW — CIVIL APPEAL — DECLARATORY ACTION — INCIDENCE OF ICMS ON TRANSMISSION (TUST) OR DISTRIBUTION (TUSD) SYSTEM USAGE TARIFFS — STJ REPETITIVE THEME NO. 986 (RESP NO. 1.699.851/TO) — LEGALITY OF THE CHARGE — APPLICABLE MODULATION OF EFFECTS — INJUNCTION IN FORCE BEFORE 03/27/2017 — DEFAULT INTEREST AND MONETARY CORRECTION AGAINST THE PUBLIC TREASURY — PREQUESTIONING — APPEAL KNOWN AND PARTIALLY GRANTED.
Structured Body (s_i)	I. CASE UNDER EXAMINATION: Appeal filed by the defendant against a sentence that declared the non-existence of a legal-tax relationship regarding the collection of ICMS on TUST and TUSD [...] II. MATTER IN DISPUTE: The possibility of charging ICMS on a calculation base that includes the Tariffs is discussed [...] III. REASONS FOR DECISION: The STJ defined, in Repetitive Theme No. 986, that TUST and TUSD integrate the base [...] IV. DISPOSITION: Civil Appeal known and partially granted.

Indexing employs the FAISS library [Johnson et al. 2019] with inner product search over previously normalized vectors, ensuring equivalence to cosine similarity. We evaluate two representation strategies for the queries q_i:

Baseline Approach: The representation derives directly from applying the embedding model to the original Structured Body:

$$v_{i,\text{base}} = f(q_i) \tag{2}$$

HyDE Approach: Grounded in the two-stage framework proposed by [Gao et al. 2023], semantic expansion decouples dense retrieval into generative relevance modeling and document-document vector matching. First, the query q_i is submitted to an instructed language model that generates a set of $M = 5$ hypothetical target documents (H_i), capturing the expected technical syntax and legal terminology:

$$H_i = \left\{h_i^{(1)}, \ldots, h_i^{(M)}\right\}, \quad \text{where } h_i^{(m)} = \text{LLM}(q_i) \tag{3}$$

Next, the contrastive embedding model encodes these synthetic documents. Crucially, the encoder's dense bottleneck acts as a lossy compressor, filtering out specific hallucinated details while preserving the core semantic manifold. This shifts the retrieval task from asymmetric query-document matching to symmetric document-document similarity in latent space. The expanded representation $v_{i,\text{hyde}}$, replacing the original query vector, is obtained via mean pooling over the generated embeddings:

$$v_{i,\text{hyde}} = \frac{1}{M} \sum_{m=1}^{M} f\left(h_i^{(m)}\right) \tag{4}$$

In the retrieval process, two strategies are implemented:

- **Dense Retrieval:** Produces a ranking $R_{i,\text{dense}}$ based on the descending order of cosine similarity between the query and the indexed vectors e_j.
- **Hybrid Retrieval:** Incorporates the BM25 algorithm over the textual collection, generating a lexical ranking $R_{i,\text{lex}}$. The system integrates the signals via Reciprocal Rank Fusion (RRF):

$$R_{i,\text{hybrid}} = \text{Fusion}(R_{i,\text{dense}},\ R_{i,\text{lex}}) \tag{5}$$

3.3. Prompt Engineering for HyDE

To generate the hypothetical documents (h_i), the query q_i was inserted into a template designed to force the LLM to mimic the technical jargon of a Brazilian appellate court. The prompt is presented below (translated to English for illustrative purposes; the original execution was strictly in Portuguese):

> *"You are an expert in the jurisprudence of the State Court of Mato Grosso do Sul (TJMS). Your task is to generate the SYLLABUS of a judicial decision based on the case description. The syllabus must EXACTLY follow the TJMS syllabus format: a sequence of KEYWORDS and THEMES in uppercase, separated by periods, without running text, without introductions, without numbering. Example format: 'CIVIL PROCEDURAL LAW. INTERNAL INTERLOCUTORY APPEAL. ADMISSIBILITY. STF THEME 339. SUFFICIENT REASONING. GROSS ERROR. APPEAL DENIED.' Respond ONLY with the syllabus, without any additional explanation. Case description: [q_i]"*

Crucially, the LLM operates strictly zero-shot, without access to the true target header (c_i). In this framework, the prompt effectively directs the model to perform a zero-shot structured abstraction and summarization task, transforming the unstructured narrative query into the expected institutional keyphrase format of the court.

3.4. Experimental Setup and Execution Environment

The experiments were conducted in a controlled environment with GPU acceleration NVIDIA RTX A6000. Semantic expansion employed the `google/gemma-3-12b-it` model [Gemma Team et al. 2024], operating with a temperature of $T = 0.7$ (to encourage lexical variability) and a limit of 200 tokens. Vector extraction was performed using the `facebook/mcontriever-msmarco` model [Izacard et al. 2022] in its native zero-shot configuration. We opted for this model because it is a robust multilingual zero-shot retriever specifically designed for unsupervised scenarios, making it an ideal baseline to evaluate the isolated contribution of HyDE without the confounding variables of task-specific fine-tuning.

Header embeddings were pre-computed and statically indexed in FAISS (`IndexFlatIP`). It is worth noting that all hypothetical syllabi were generated and

statically stored prior to the retrieval stage, ensuring the reproducibility of the experiment without suffering dynamic variations from the LLM's stochasticity.

3.5. Evaluation Protocol and Metrics

We adopt a controlled known-item retrieval protocol. Since each query q_i originates from an appellate decision, only its respective Header c_i is strictly relevant. Thus, the ideal relevance function $\text{rel}(q_i, c_j)$ equals 1 if $i = j$, and 0 otherwise.

Let rank_i be the ordinal position of the correct document. System performance is measured by three classic Information Retrieval metrics:

- **Recall@k:** Measures the proportion of queries where the relevant document was retrieved within the top k positions.
- **Mean Reciprocal Rank (MRR):** Computes the average of the reciprocal ranks, rewarding items positioned at the absolute top.
- **Normalized Discounted Cumulative Gain (nDCG@k):** Evaluates document position within a k cutoff using a logarithmic penalty. Given our single-document scenario, the Ideal DCG (IDCG) is fixed at 1.

The experiment compares these metrics across four scenarios: (i) Dense Baseline; (ii) Dense HyDE; (iii) Hybrid Baseline; and (iv) Hybrid HyDE. All other framework components remain fixed, isolating the impact of each technique.

4. Results and Discussion

This section discusses the empirical results obtained on the TJMS corpus, organized around the three Research Questions (RQs).

4.1. Overall Effectiveness Analysis (RQ1)

For RQ1, we evaluated three isolated retrieval strategies: (i) **BM25 (Lexical)**, a sparse baseline focused on exact term matching; (ii) **Embedding (Dense)**, a semantic baseline utilizing the original query vector without expansion; and (iii) **HyDE (Expanded)**, the proposed approach that replaces the original query with the embedding of the LLM-generated hypothetical document. Table 2 consolidates the metrics for cutoffs $k \in \{1, 5, 10\}$.

Table 2. Overall Effectiveness Results in Syllabus Retrieval (RQ1)

Strategy	R@1	R@5	R@10	MRR	nDCG@5	nDCG@10
BM25 (Lexical)	0.531	0.629	0.693	0.574	0.629	0.693
Embedding (Dense)	0.086	0.192	0.268	0.131	0.139	0.163
HyDE (Expanded)	0.345	0.545	0.619	0.427	0.448	0.473

Table 2 reveals two fundamental dynamics of the domain. First, the impact of HyDE is remarkable in the dense paradigm: the MRR jumped from 0.131 (pure vector) to 0.427, and top-1 precision quadrupled (with R@1 rising from 0.086 to 0.345). This quantitative gain confirms that generating a hypothetical document considerably reduces the vocabulary mismatch inherent to zero-shot dense approaches.

However, BM25 (Lexical) outperformed all purely semantic approaches, highlighting the resilience of sparse algorithms. This phenomenon reflects the nature of the

legal corpus: the Structured Body frequently contains unambiguous textual keys (e.g., case numbers, statute articles, or specific values) that appear identically in the Header. The sparse algorithm captures these exact intersections optimally, whereas dense models tend to dilute highly specific keys in favor of a broader contextual representation.

To verify the statistical significance and magnitude of HyDE's gain over the direct vector model, a paired Wilcoxon signed-rank test was applied to the reciprocal rank distributions ($\alpha = 0.05$), complemented by effect size estimation ($r = |Z|/\sqrt{N}$). The test rejected the null hypothesis ($p < 0.001$) with a large effect size ($r = \mathbf{0.52}$). This confirms that the substantial gain (MRR rising from 0.131 to 0.427) reflects a large practical effect rather than a statistical artifact of the sample size ($N = 501$).

In summary, in response to RQ1, the empirical results confirm that generation-guided semantic expansion significantly improves the effectiveness of standalone dense retrieval, successfully bridging the vocabulary gap between narrative queries and specialized legal headers.

4.2. Impact of Hybrid Retrieval (RQ2)

For RQ2, we investigated whether fusing semantic and exact lexical signals generates a synergistic effect superior to isolated models. Two combined configurations via the Reciprocal Rank Fusion (RRF) algorithm are introduced: Hybrid (Dense + BM25), which merges the vector baseline with pure lexical scores; and Hybrid (HyDE + BM25), which integrates the expanded vector of the hypothetical document with the exactness of the sparse algorithm. Table 3 presents the quantitative results exclusively for these fused strategies.

Table 3. Exclusive Performance Comparison of Hybrid Configurations (RQ2)

Strategy	R@1	R@5	R@10	MRR	nDCG@10
Hybrid (Dense + BM25)	0.176	0.499	0.649	0.307	0.388
Hybrid (HyDE + BM25)	0.483	0.675	0.733	0.571	0.610

Hybridization revealed a powerful synergistic effect. Contrasting Tables 2 and 3, the pure lexical approach maintains the upper bound of effectiveness in this corpus, whereas the Hybrid (HyDE + BM25) configuration substantially outperforms all vector-based architectures. Fusing the expanded vector with BM25 raised the Recall@1 from 0.345 (isolated HyDE) to 0.483, reaching a technical tie in the MRR metric with the lexical baseline (0.571 versus 0.574).

In the legal domain, appellate decisions frequently share a macro-semantic context while differentiating themselves through highly specific data (e.g., exact acronyms such as "ICMS" and "TUST" or mentions of "Repetitive Themes"). In this scenario, while the HyDE component ensures conceptual alignment, the lexical component anchors the search on critical literal matches. This joint action prevents the correct document from suffering severe logarithmic ranking penalties in nDCG@10, correcting the main weakness of isolated dense retrieval.

To answer RQ2, our evaluation demonstrates that while the lexical BM25 baseline remains highly competitive in Portuguese legal texts, the hybrid approach (HyDE + BM25) closes the performance gap, achieving comparable effectiveness and

establishing a robust alternative for vector-based systems. Paired Wilcoxon tests confirm that Hybrid (HyDE + BM25) significantly outperforms standalone HyDE ($p < 0.001$, $r = \mathbf{0.46}$), while showing no statistically significant difference compared to pure BM25 ($p = \mathbf{0.741}$, $r = \mathbf{0.015}$), formalizing their statistical equivalence.

4.3. Influence of Query Granularity (RQ3)

The motivation for this experiment is to optimize the cost-benefit ratio in practical search systems. Since language models have context-window constraints and per-token processing costs, identifying the minimum amount of information required to generate an accurate document is essential to the solution's viability.

To answer RQ3, we mapped how query length influences generation in HyDE by isolating three granularity scenarios from the Structured Body: (i) **Short Scenario**: factual text from the *I. CASE UNDER EXAMINATION* section; (ii) **Medium Scenario**: combines the previous section with *II. MATTER IN DISPUTE*; and (iii) **Long Scenario**: the full body of the syllabus (facts, controversy, reasons for decision, and disposition). Table 4 consolidates the quantitative impacts of these variations.

Table 4. Impact of Query Length on HyDE Performance (RQ3)

Scenario	Avg. Length	HyDE (Expanded)		Hybrid (HyDE + BM25)	
	(words)	R@1	MRR	R@1	MRR
Short	∼118	0.263	0.347	0.395	0.481
Medium	∼180	0.329	0.412	0.455	0.545
Long	∼656	0.339	0.425	0.477	0.566

The data indicate that providing more context increases effectiveness, but performance stabilizes rapidly at larger text volumes. The most significant leap occurs from the Short to the Medium Scenario. This demonstrates that explicitly adding the legal controversy provides the necessary semantic anchoring to infer the main thesis, an outcome unattainable relying solely on the factual narrative.

More importantly, this reveals a practical efficiency limit for real-world systems: the Medium Scenario consolidates itself as the optimal operating point, delivering substantial gains with a lean informational load (∼180 words). The Long Scenario, despite reaching the technique's upper bound of precision, requires a textual volume nearly four times larger (∼656 words). This severe increase in computational cost and latency is not justified by the strictly marginal gain, suggesting that the final sections of the appellate decision yield diminishing returns for hypothetical document generation.

Addressing RQ3, the sensitivity analysis indicates that medium-length queries, containing both facts and controversy, represent the optimal operational point, maximizing precision while avoiding the computational overhead of processing overly extensive full-text documents. Wilcoxon tests confirm that the improvement from Short to Medium granularity is statistically significant ($p < 0.001$, $r = \mathbf{0.21}$), whereas the gain from Medium to Long shows no statistically significant difference ($p = \mathbf{0.302}$, $r = \mathbf{0.05}$), corroborating the Medium scenario as the saturation point.

4.4. Discussion and Limitations

BM25's strong performance reflects protocol boundaries: deriving queries and targets from the same decision inflates lexical overlap (e.g., acronyms such as CDA and IPTU), overestimating exact matching versus exploratory search. The hybrid architecture (HyDE + BM25) bridges this semantic gap while preserving literal matching. Qualitatively, query inspection reveals three dynamics: (1) HyDE overcomes severe vocabulary mismatch by injecting institutional jargon (e.g., "CNJ Resolution 547/2024", "STF Theme 1.184"), elevating targets from rank >10 to 1; (2) hybrid fusion resolves cases where HyDE captures the doctrinal frame but cites no statute (rank 6), requiring BM25's literal matching (rank 8) to reach rank 1; and (3) fabricated precedents ("STJ Theme 1.178") and dropped discriminative terms ("usucapião") displace dense vectors, moving targets from rank 1 to >10 (7 of 501 queries; one degradation), the dominant failure being over-abstraction into generic syllabi.

This study presents specific limitations. First, the 501-decision corpus lacks production scale and distractor density among overlapping precedents. Second, deriving queries from narrative bodies omits the broader vocabulary gap present in independently formulated queries (e.g., petitions). Third, restricting relevance to a single paired target penalizes template-based near-duplicates as false positives, establishing a lower bound on measured performance. Fourth, prompts targeted full synthetic headers matching the TJMS structure, leaving simpler prompts (e.g., keywords or statutory lists) and traditional pseudo-relevance feedback for future ablation studies. Finally, validation relied on a single dense encoder and a 12B LLM, omitting latency-effectiveness trade-offs of smaller language models (SLMs).

5. Conclusion

This work investigated the impact of generation-guided semantic expansion (HyDE) on the retrieval of Brazilian legal precedents. By addressing our research questions using a known-item retrieval protocol on a real-world TJMS corpus, we demonstrated that HyDE significantly improves the effectiveness of standalone dense retrieval (RQ1). Furthermore, we established that hybridization (HyDE + BM25 via RRF) achieves overall performance comparable to the strong lexical baseline (RQ2). Finally, our sensitivity analysis revealed that medium-length queries integrating facts and controversy offer the optimal balance between computational cost and precision gain, effectively preventing processing waste (RQ3).

A primary contribution of this study is providing a controlled empirical evaluation of HyDE in national case law. Building upon the identified limitations, future work should scale the evaluation to significantly larger corpora with higher distractor density, conduct prompt ablation studies comparing full synthetic headers against simpler keyword expansions, evaluate smaller generative models (SLMs) to assess latency-precision trade-offs against traditional query expansion baselines, validate against independently formulated queries, and explore domain-adapted embedding models.

In summary, the results confirm the viability of hybrid generative architectures not only to bridge the vocabulary gap in legal research but also to pave the way for

more transparent, responsive, and interpretable legal search systems.

Acknowledgments

This work received financial support from FUNDECT (TO 2102/2025, Process 83/046.532/2025) through a Scientific Initiation (PIBIC) scholarship, with additional support from CNPq, CAPES, and UFMS.

References

Ariai, F., Mackenzie, J., and Demartini, G. (2025). Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*, 58(6):1–37.

Bueno, T. C. D., von Wangenheim, C. G., de Silva Mattos, E., Hoeschl, H. C., and Barcia, R. M. (1999). Juris Consulto: Retrieval in jurisprudencial text bases using juridical terminology. In *Proceedings of the 7th International Conference on Artificial Intelligence and Law (ICAIL 1999)*, pages 147–155, Oslo, Norway. ACM.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1–50.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., and Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330. Association for Computational Linguistics.

Conselho Nacional de Justiça (2024). Justiça em números 2024: ano-base 2023. Technical report, Conselho Nacional de Justiça (CNJ), Brasília, DF.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, Boston, Massachusetts, USA. ACM.

Fernandes, L. C., Ribeiro, L. d. S., de Castro, M. V. B., Pacheco, L. A. d. S., and Sandes, E. F. d. O. (2026). JurisTCU: A Brazilian Portuguese information retrieval dataset with query relevance judgments. *Language Resources and Evaluation*, 60(1):23.

Gao, L., Ma, X., Lin, J., and Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777. Association for Computational Linguistics.

Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, 5:181–196.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 9459–9474, Vancouver, Canada.

Lin, J., Nogueira, R., and Yates, A. (2022). *Pretrained transformers for text ranking: BERT and beyond*. Springer Nature.

Luz de Araujo, P. H., de Campos, T. E., Braz, F. A., and da Silva, N. C. (2020). VICTOR: A dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 1449–1458, Marseille, France. European Language Resources Association (ELRA).

Oliveira, L. S., de Castro, F. F., and Rocha, E. S. S. (2024). Recuperação da informação jurídica: um mapeamento sistemático. In *Anais do XXIV Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ENANCIB)*, pages 1–19, Vitória, ES. ANCIB.

Rabelo, J., Goebel, R., Kim, M.-Y., Kano, Y., Yoshioka, M., and Satoh, K. (2022). Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.